

Direkte Rede zur Analyse von Figurenbeziehungen

Markus Krug, Albin Zehe

Überblick

- Ziele in Kallimachos II
 - ATHEN
 - Szenenübergänge
 - Figuren(-beziehungen)
 - Analyse direkter Rede

Ziele in Kallimachos II

- Erkennung von Szenenübergängen
- Charakterisierung von Figuren(-beziehungen)
- Veränderungen über den Verlauf eines Romans
- ATHEN

Neues von ATHEN (I)

- Webathen, aktuell in Entwicklung mit dem Ziel einer reibungslosen Integration in eine end-to-end pipeline

Seit diesem Tage lastete ein erneuter Druck auf den **Bewohnern** des Hofes. **Tralgoth** lauerte auf alles, was in **seiner** Nähe vorging. Halbe Nächte lang saß **er** im Bette wach und horchte auf **Hendriks** Schritte, **der** oben in **seiner** Stube auf und nieder ging. Das Nachtlicht brannte grell bis zum Morgengrauen. Dann begann der Tag mit seiner düsteren Monotonie.



- Webapp, mit dem Ziel so viel wie möglich unseres Textminings im Frontend laufen zu lassen

Neues von ATHEN (II)

- Aktuelle Features:
 - Laden und Speichern von (JSON, XML, XMI, TXT)-Dokumenten
 - Annotieren beliebiger Typen
 - NLP (-WIP-):
 - Tokenisieren
 - POS-Tagging
 - Parsen
 - NER
 - Coreference
 - Relationserkennung
 - Z.B. mittels Deeplearn.js („Web-Tensorflow“)
 - Ontologiebasierte Informationsextraktion (Der eigentliche Grund für die Existenz)

Neues von ATHEN (II)

- Hauptziel für Kallimachos:
 - Entwurf einer Gesamtpipeline vom Scan bis zur Analyse von Dokumenten
- Aktuelle Version unter:
 - webathen.informatik.uni-wuerzburg.de

Erkennung von Szenenübergängen

- Annotationen schwieriger als erwartet
 - Wie definieren sich Szenen?
 - Wann findet ein Szenenübergang statt?
- Bisher:
 - Versuch eines einheitlichen Annotationsschemas
- Zukunft:
 - Formalisierung des Annotationsschemas
 - Umsetzen in Programmcode

Erkennung von Szenenübergängen

- Was ist eine Szene?
- Möglichkeit 1

Eine Szene ist ein Textausschnitt, in dem erzählte Zeit und reale Zeit übereinstimmen

- ➔ Erzeugt eine lineare Anordnung von Szenen
- ➔ Erste Experimente der Annotation haben keine gute Übereinstimmung der Annotatoren ergeben


Erkennung von Szenenübergängen

- Was ist eine Szene?
- Möglichkeit 2

Ein Szenenwechsel findet bei einem Wechsel des Ortes, der Zeit, oder der vorhandenen Figuren statt.

- ➔ Was genau ist ein Ortswechsel? („Vom Wohnzimmer in die Küche?“)
- ➔ Was wenn die Figur anwesend war, aber nicht aktiv?
- ➔ Sind Szenen überhaupt linear oder gar hierarchisch angeordnet?

Charakterisierung von Figuren(-beziehungen)

- Wichtiger Indikator für Figurenbeziehungen:
 - Dialoge
- Letztes Jahr:
 - Erkennung von Sprechern und Angesprochenen bei bekannten direkten Reden
 - Sentiment Analysis 
- Jetzt:
 - Automatische Erkennung von Redepassagen
- Plan:
 - Genauere Charakterisierung einzelner Figuren

Hilf Himmel!
rief sie ihm entgegen,

[DS → Speaker = P1](,) SAY P1

Sprechererkennung ✓

- Regelbasiertes System
- Erkennt mit hoher Trefferquote (82,2%) den Sprecher
- Erkennt mit relativ hoher Quote (66,8%) den Angesprochenen

Hilf Himmel!

rief sie ihm entgegen,

417 794

[DS → Speaker = P1](,) SAY P1

Sentiment Analysis

- Erkennung der Polarität eines Satzes

Ich bin froh ↔ Ich bin traurig

- Häufige Probleme:

- Negationen Ich bin nicht froh
- Sarkasmus Na, das ist ja großartig...
- ...



Sentiment Analysis

- Größtes Problem:
 - Fehlende Annotationen für deutsche Romane
- Dennoch erste vielversprechende Ergebnisse
 - 66% für binäre Klassifikation (positiv/negativ) mit SVM
 - Interpretierbare Featuregewichte
 - SVM lernt sinnvolles Modell, nur zu wenige Daten um wirklich gut zu werden



Automatische Erkennung direkter Rede

- Nichttriviales Problem:
 - Inkonsistente Verwendung von Anführungszeichen
 - Verschachtelte direkte Reden
 - Teile von Romanen im „Dramenstil“
 - Anführungszeichen nur bedingt geeignet!
 - Verwende Machine Learning!

Direkte Rede - Trainingsdaten

- Möglichst geringer manueller Aufwand:
Verwendung von „schwachen“ Labels
- Automatische Extraktion von Labels aus
aufbereitetem Kernkorpus
→ Alles innerhalb von Anführungszeichen ist
direkte Rede
- Fehlerrate von $\sim 3\%$
- $\sim 36\%$ direkte Rede

Direkte Rede - Testdaten

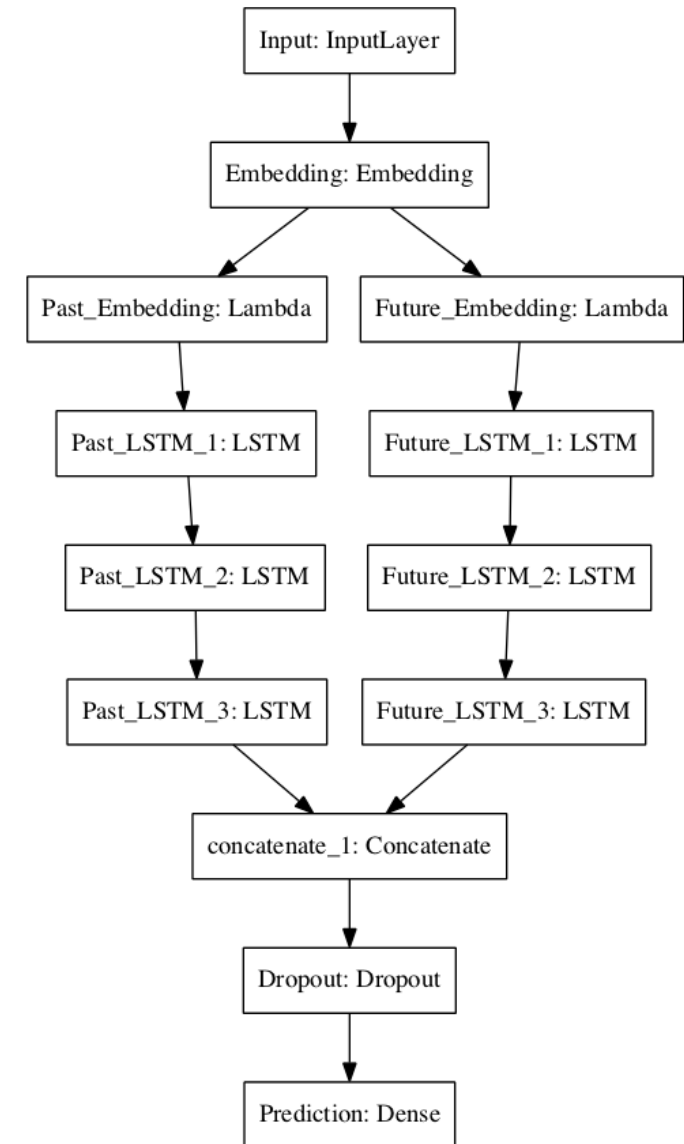
- Zum Testen manuell annotierte Daten nötig
- 50 Textausschnitte aus Schemaliteratur
- Relativ wenig direkte Rede (~18%)

Direkte Rede - Task

- Zuerst: Satzklassifikation
 - Enthält ein Satz mindestens einen Teil direkter Rede?
 - Erkennungsrate bereits auf menschlichem Niveau (~84%)!
- Daher: Wortklassifikation
 - Klassifiziere jedes Wort als innerhalb oder außerhalb einer direkten Rede
 - Genauere Klassifikation → Anspruchsvoller

Direkte Rede - Modell

- Rekurrentes neuronales Netz
- Einbeziehen von Kontext vor und nach dem Zielwort



Direkte Rede - Ergebnisse

- Training auf (ggf. Teil der) schwachen Labels des Kernkorpus
- Auswertung auf
 - Verbleibenden schwachen Labels → 83% Treffer
 - Manuellen Labels → 90% Treffer

Direkte Rede - Pläne

- Verfeinerung des Netzes
- Featureanalyse!
 - Welche Wörter sind wichtige Signale?

DANKE!